R E P O R T R E S U M E S

ED 015 645 64 EM 004 913

RECOMMENDATIONS FOR REPORTING THE EFFECTIVENESS OF PROGRAMMED INSTRUCTION MATERIALS.

AMERICAN EDUCATIONAL RESEARCH ASSN., WASH., D.C.

REPORT NUMBER NDEA-VIIB-219 PUB DATE OCT 65

AMERICAN PSYCHOLOGICAL ASSN., WASHINGTON, D.C.

REPORT NUMBER DR-5-1013

NATIONAL EDUCATION ASSN., WASHINGTON, D.C.

CONTRACT OEC-SAE-9538

EDRS PRICE MF-$0.25 HC-$1.12 26P.

DESCRIPTORS- *PROGRAMED INSTRUCTION, *RESEARCH UTILIZATION, *PROGRAM EFFECTIVENESS, *PROGRAM EVALUATION, *PROGRAM GUIDES

BASIC PREMISE OF THIS REPORT IS THAT INSTRUCTIONAL EFFECTIVENESS MUST BE JUDGED FOR EACH PROGRAM ACCORDING TO ITS DEMONSTRATED MERITS. GENERAL AND SPECIFIC RECOMMENDATIONS ARE LISTED FOR POTENTIAL PROGRAM USERS AND PUBLISHERS. SUPPLEMENTS CONTAIN INFORMATION ON PROGRAM MANUALS AND TECHNICAL DOCUMENTATION. THIS REPORT WAS PREPARED BY THE JOINT COMMITTEE ON PROGRAMMED INSTRUCTION AND TEACHING MACHINES OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, THE AMERICAN PSYCHOLOGICAL ASSOCIATION, AND DEPARTMENT OF AUDIO-VISUAL INSTRUCTION (NATIONAL EDUCATION ASSOCIATION). (LM)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED015645

RECOMMENDATIONS FOR REPORTING THE EFFECTIVENESS OF PROGRAMMED-INSTRUCTION MATERIALS

Prepared

by the

Joint Committee on Programmed Instruction and Teaching Machines
(A.E.R.A. -- A.P.A. -- D.A.V.I./NEA)

September, 1964
(Revised October, 1965)

I. INTRODUCTION

Background

This report has been prepared by the Joint Committee on Programmed Instruction and Teaching Machines of the American Educational Research Association, the American Psychological Association, and Department of Audio-Visual Instruction (National Education Association).[1] The work of this committee has been supported by the Educational Media Branch, Office of Education, U.S. Department of Health, Education, & Welfare, under Title VII of the National Defense Education Act. The parent associations have charged the committee with providing useful guidance to publishers and purchasers of programmed-instruction materials.[2] The present recommendations are intended to help in improving the effectiveness of program selection and utilization. This report supplements the 1962-63 committee report, to which the reader is referred for background.[3]

1, 2, 3: See Footnotes at end of this report

EM 004 015

Purpose and Content

The report provides assistance primarily to potential users concerned with the selection and effective use of instructional programs. It also provides guidance to those who publish programs or report data on program effectiveness. The present report deals with recommendations concerning information on the effects that a given program can be shown to produce, regardless of how these effects may relate to the user's purposes. Supplement I to this report contains suggestions for information to be included in a Program Manual for teachers and other users who require information about program characteristics. Supplement II contains recommendations intended to serve as a guide for those who are preparing technical documentation in support of statements about the outcomes that a program can produce.

II. RECOMMENDATIONS ON REPORTED EFFECTIVENESS OF PROGRAMS

The basic premise of this report is that instructional effectiveness must be judged for each program according to its demonstrated merits. Evidence for the effectiveness of a program should be based on a detailed study which has been fully documented in a technical report. As emphasized in its 1962-63 report, the committee takes the position that effectiveness of each program must be determined by measurement of the instructional outcomes which that program's use can be shown to bring about. At the present state of the art, users generally cannot assess the effectiveness of a particular program reliably by mere inspection of the program or by reference to statements about its developmental history.

A. General Recommendations

Recommendation 1:

Evidence for the effectiveness of a program should be based on a carefully conducted study which shows what the program's use accomplished under specified conditions. Such a study must employ suitable before and after measurements, with control procedures to insure that

effects attributed to the program can be clearly distinguished from the effects of other instruction.

Recommendation 2:

The results of the evaluation study should be carefully documented in a technical report prepared in keeping with accepted standards for scientific reporting. (Specific recommendations for the preparation of such documents are presented in Supplement II to this report).

Recommendation 3:

All claims or statements about the effectiveness of a program should be supported by specific reference to the evidence contained in the technical report.

B. More Specific Recommendations

It is assumed that data on the effectiveness of programs will be obtained and reported by: (1) program producers, (2) using agencies, including school systems, and (3) projects conducted by universities and other research agencies. Accordingly, some further specific recommendations and suggestions, given below, are addressed to prospective users, program publishers, reviewers, and research agencies or institutions which conduct or report assessment studies.

The uses of program-assessment data differ depending upon the needs and technical experience of the user. For most teachers and school administrators, reports are needed which report the effectiveness of an instructional program in fairly straightforward terms that are quickly comprehensible without examining detailed technical data. On the other hand, a detailed account of all experimental procedures and instruments used in assessment is needed for the technical evaluator who must critically analyze the study to see if the summarized results and interpretations are warranted.

1. Recommendations for the prospective purchaser or user:

a. Prospective users should evaluate each program on its own merits according to its demonstrated effectiveness in producing specified outcomes.

b. In determining the suitability of any program for a particular purpose, the prospective user should first formulate his own objectives in as much detail as possible and then evaluate the program in relation to these objectives in the light of three things:

(1) The apparent appropriateness of the program content for his purposes, as based on inspection of the program itself and of the producer's statement of the program's objectives. These objectives may be inferred from tests supplied by the producer for measuring the intended outcomes of the program.

(2) Consideration of factors affecting practicality, or feasibility of use, such as the unit cost of the program, initial and maintenance cost of a machine (if required), and factors affecting supervision, scheduling, and other aspects of administration.

(3) Evidence on the demonstrable effectiveness of the program in terms of outcomes relevant to the user's objectives. (These may include motivational, or attitudinal effects, as well as subject-matter competences.)

c. The prospective user is advised to ignore all claims for the effectiveness of a program which are not backed up by appropriate data that have been subjected to competent evaluation. Advice on the soundness of claims for program effectiveness should preferably be obtained from a technical advisor or reviewer who has competence in the fields of educational psychology, measurement, and experimental design, and who has reviewed available reports on the effects of the program in the light of technical recommendations identified in Section 5, below.

d. In addition to consulting reviews published in professional journals, users should seek all available data on demonstrated performance characteristics of the program, not only from information supplied by the producer, but also from reports prepared independently: for example, reports prepared by school systems, research projects, or other agencies that have conducted program-assessment studies of the particular program.

2. Recommendations for program publishers:

The following recommendations, plus recommendations 4a-4e below, are offered to assist the program producer in providing necessary information which will help users make intelligent choices among available programs.

a. The publisher should state in detail the minimum objectives of his program, preferably in terms of specific behaviors or competences which its use is intended to achieve for specified kinds of learners.

b. Publishers should refrain from promoting a program in terms of general statements about the value of programmed instruction as a general "method," or on the basis of statements about its effectiveness not supported by detailed data, as recommended above.

c. Publishers should provide a program manual, preferably one that can be updated or supplemented as new data on the program become available. (See suggestions for the context of such a manual provided in Supplement I to this report.)

d. Preliminary limited editions of programs, prior to validation by definitive evaluative studies meeting the conditions of technical adequacy indicated herein, should be issued to facilitate collection of evaluation data. These should always be clearly identified to the purchaser as experimental or preliminary editions.

e. Publishers should use a suitably descriptive title for the program which appropriately delimits the scope of the subject matter and skills taught. Relatively longer titles and use of sub-titles and detailed tables of contents are recommended.

3. Recommendations for reviewers:

To assist users in evaluating programs, those who prepare reviews might, in addition to expressing their opinions about the suitability of the program content and objectives, be guided by the following suggestions and by recommendations 4a-4d below.

a. Take into account all available assessment data.

b. Evaluate and interpret such data in the context of technical considerations such as those set forth below and amplified in Supplement II to this report.

c. Make available (for example, in a supplementary report or by deposit with the American Documentation Institute) any relevant details of his analysis of assessment data which require more space than is appropriate for a published review in professional journals.

d. Utilize a procedure and format of reporting which provides a thorough analysis required of the program. (When appropriate, reviewers might consider employing a set of topical headings related to program appropriateness, such as producer's statement of program objectives, appropriateness of objectives to current curricular concepts, suitability of objectives and program to the designated student population, etc. Where data from a formal study to assess program effects are available, the reviewer should evaluate the data in the light of the criterion measures used, adequacy of description of test populations, description of conditions of experimental and intended program use, degree of correspondence between conditions of testing and of intended application, etc.).

4. Recommendations addressed jointly to program producers, reviewers, and technical advisors.

The following recommendations recognize that data on the effects of programs may vary from impressions based on observations of one or two subjects, as they work through a program, to a full-scale, formal study in which the program's specific effects on learning, retention, motivation, and application of knowledges and skills are determined for representative populations of students under varying conditions of use. In the formal study, these data may be analyzed to show differing effects for sub-groups of varying aptitude and background.

Informal tryouts and subjective impressions can be useful when intended to serve as guidance to the programmer in revising early versions of a program. The teacher also receives some value from informal tryout when a rough, overall "screening test" is desired to help decide whether or not a program seems to "work" (in the sense of being generally suitable for its intended purpose.[4])

The recommendations that follow are concerned primarily with the reporting of formal and rigorous assessment studies which are required for determining in some detail

the performance characteristics of a program -- that is, the specific outcome which a program can be shown to be reliably capable of producing.

a. Reported data on effectiveness should refer to the effects produced by the program itself, unless other instructional sources are clearly identified, and their contribution is assembly. (Although most schools use programs in conjunction with other media of instruction, it will generally help a prospec;tive user to know what the program alone actually contributes to the student's knowledge or proficiency, in addition to what is contributed by other elements in the instructional situation.)

b. Program producers should cite the available evidence their own studies of the program to document any claims they make about the effectiveness of the program. Publishers as well as reviewers should also cite any pertinent evidence available from all other documented studies of the program that are know to them.

c. Publishers and reviewers should differentiate clearly and explicitly between (1) mere opinions, of experts or others, about the probable effectiveness of the program, as distinguished from (2) documented evidence on the outcomes its use has been actually shown to produce.

d. While brief summary statements may often appear in advertising copy or brochures, or on the cover or label of a program, or in a program manual, or a review, such summary statements should always cite the tehcnical report on which they are based, so that the correspondence between interpretive statemenrs and underlying data is made explicit.

e. Summaries of information concerning a program's demonstrated effectiveness should be made widely available in published sources of general distribution. It is suggested that program producers report such data in a program manual, and that using agencies (such as school systems and research projects) also publish the results that they obtain, on any commercially available program, in appropriate professional journals. These agencies should provide copies of their reports to the publisher of the program as soon as they are completed and ready for publication.

f. To insure continued availability of techincal reports and of data not published in full in standard books or journals, at least one complete copy of the technical report and of all basic data tabulations should be furnished to a suitable depository such as the American Documentation Institute or University Microfilms, and this fact should be noted in the program manual.

5. Recommendations Concerning Technical Reporting.

These recommendations, in addition to Recommendations 4a-4f above, are addressed to those concerned with obtaining, reporting, or evaluating the adequacy of information from empirical studies of the effects of programs. The recommendations summarized here are amplified in Supplement II to this report.[5]

a. In accordance with basic criteria of scientific reporting, the entire evaluation procedure should be reproducible. This applies to the derivation, administration, and description of criterion measures as well as to the selection of the experimental design and procedure. The technical report should describe the procedures used and the results obtained in such a way that a technically-qualified person (1) can assess the validity of the statements concerning what outcomes the program's use will achieve, and (2) could replicate the study in substantially identical fashion.

b. To satisfy this basic requirement, the technical report should give full details on all relevant aspects of the evaluation study, including criterion measures, characteristics of subjects, conditions of program use and data collection, experimental design, and data obtained. Some of the more specific aspects of such topics dealt with in Supplement II are:

(1) procedures employed in measuring of retention, transfer, and attitude, and other dependent measures such as time to reach a criterion;

(2) characteristics of students as indicated by measures of prior knowledge, competence, intelligence, aptitude, etc.

(3) procedures and scheduling of program use, related instruction, conditions of administration;

(4) procedures used in sampling and assignment of subjects, use of control and comparison groups, controls for so-called "Hawthorne" effect, and related spurious influences.

(5) Processing, tabulations, analysis, and summarization of data, tests of significance and reporting of fiducial limits, etc.

FOOTNOTES

(1) Committee members are: Harry F. Silberman, Evan R. Keislar, Robert Glaser, and Arthur A. Lumsdaine, Chairman (AERA); Richard S. Crutchfield, James G. Holland, and Lawrence M. Stolurow (APA); and Jack V. Edling, Edward B. Fry, Wesley C. Meierhenry, and Paul R. Wendt (DAVI). Ernst Z. Rothkopf served as consultant, and Brett B. Hamilton as staff assistant, in the preparation of this report. Helpful contributions were made to the preparation of the present statement by members of a cooperating committee of the American Society for Training and Development (formerly American Society of Training Directors) under the chairmanship of Leonard C. Silvern. The present report represents a concensus of the Joint Committee members rather than official policy of AERA, APA, or DAVI. Further suggestions from program writers, publishers, or users are invited.

(2) A useful guide to available programs for school subjects is the government publication entitled Programs '63, prepared by the Center for Programed Instruction (U.S. Office of Education, Publication No. OE-34015-63; 814 pp., $2.50). This publication lists some 350 programs reported to be commercially available by the end of 1963, and includes descriptive information, price, and one or more sample sequences from each program, though with no attempt to evaluate the programs. (See also the later similiar compilation, Programs '65). Another useful compilation of programs, which includes more programs but gives less detail on each program, is Programed Learning: A Bibliography of Programs and Presentation Devices, edited by Carl H. Hendershot, and published by Delta College, University Center, Michigan. This listing has been updated quarterly (L/C Cat. No. 64-11824; $3.50).

(3) The 1962-63 report was entitled "Criteria for Assessing Programmed Instructional Materials," and was published in Audiovisual Instruction, February, 1963, pp. 84-89; it has also been reprinted in several other educational journals and books.

(4) Suggestions concerning such classroom tryouts by the teacher are offered in the booklet, Selection and Use of Programed Materials: A Handbook for Teachers, published by the Division of Audiovisual Instruction Services of the National Educational Association, Washington, D.C. (1964; Library of Congress catalogue number 64-23523); 50¢ per copy.

(5) For further background on rationale, techniques, and problems in assessing the effectiveness of instructional programs, the reader may wish to consult the chapter by A. A. Lumsdaine entitled "Assessing the Effectiveness of Instructional Programs" in the book Teaching Machines and Programed Learning II: Data and Directions (Ed. by R. Glaser; Washington, D.C.: National Education Association, 1965).

SUPPLEMENT I

to

RECOMMENDATIONS FOR REPORTING THE EFFECTIVENESS OF PROGRAMMED-INSTRUCTION MATERIALS[1]

prepared by the AERA--APA--DAVI Joint Committee on Programmed Instruction and Teaching Machines[2]

Recommendations Concerning Program Manuals

(Revised 31 October 1965)

It is assumed that most publishers will prepare program manuals to accompany published programs. Such a manual would be used by teachers, curriculum supervisors, or others who wish information about the nature of the program. This Supplement is intended to help program producers prepare an effective manual to accompany a specific instructional program.

There are many agencies, such as large school systems, that may also wish to prepare a program manual as either a substitute for or amplification of the one furnished by the program producer. This Supplement therefore is addressed to all persons who may have occasion to prepare a program manual for use with an instructional program.

Purpose of a Program Manual

In some respects the function of a program manual is similar to that of manuals supplied with psychological or educational tests. It should provide information which will help a prospective user make appropriate and effective choices of programs. In addition, the manual can provide guidance as to the most effective application of the program.

Suggestions Concerning Contents of Program Manuals

The manual should include the following types of information:

1. Content of instructional program.
2. Description of intended student population and tryout population for which test data are reported.

---

[1]This report is a supplement to a more extensive document, "Recommendations for Reporting the Effectiveness of Programmed Instruction Materials" (Revised July, 1965); the latter is referred to herein simply as the "basic report."

[2]The membership and sponsorship of the committee is given on p.1 of the basic report.

3. Rationale for tests used to assess instructional program.
4. Evidence of effectiveness.
5. Practicality (cost, etc.).
6. Procedures for introduction and use.

General Features of a Program Manual

The program manual can be prepared so that new data on the program's effectiveness can be readily included in subsequent revisions of the manual. Each edition of a manual should be dated. Manuals which are prepared for preliminary, limited editions of programs, prior to validation by definitive evaluative studies, should be clearly identified to the purchaser as such.

1. Information concerning content of instructional program

To help the user decide whether the program content is appropriate for his purpose, the Program manual should state the minimum objectives of the program in detail, preferably by specifying student behaviors or competencies which its use is intended to achieve. The objectives should be exemplified by test items which are regarded as suitable for measuring the intended outcome of the program. These test items should be drawn from a test included or described in the manual.

It is helpful and appropriate to state objectives so that they may be compared with the newer as well as more conventional curricula. The manual might well cite comments from reviews in professional journals which indicate how a certain content is related to specific curriculum objectives. An outline of the specific content covered is often desirable.

2. Description of intended student population

The kind of student for whom the program is designed should be specified. Information will probably be included regarding grade level, cultural background, age, and prerequisite skills. Minimum competencies necessary for success in the program might be presented by prerequisite scores on a number of tests. These tests might include scholastic aptitude, reading, and specific measures of competence in the subject matter field for which the program is designed.

The manual should make clear the nature of the population used in evaluations of the program. Any cultural or sociological differences between the intended population and the tryout population should be specified.

3. Rationale for tests used to assess instructional program

The manual should give a brief description of the rationale for the selection of the criterion test for the program. It should make explicit the relation of the statement of objectives to the test used in measuring the program's effectiveness. Such clarification will help in the preparation of additional tests if desired and will aid the program-user to interpret existing data.

The manual should also indicate which of the desirable outcomes in the program's content area are not being developed by the program and thus provide a better understanding of what are reasonable expectations for the program. Such an explanation would demonstrate how various classes of test items were related to the kind of behavioral changes that reflect the objectives of the program. It is necessary not only to supply test questions to clarify the program's objectives but to indicate what kinds of answers are acceptable or nonacceptable.

Where test items used to illustrate an objective are different from those actually used in the test adopted to measure program effectiveness (as reported in the manual), such differences should be explicitly stated. Whenever test items are used for illustrative purposes, the reader should be able to find the tests from which these items were drawn. Sometimes these tests will be reproduced in full in the manual, but if not, references to the appropriate documents should be given so they may be readily located.

4. Information about the program's effectiveness

Evidence for the effectiveness of a program should be based on carefully conducted studies which show what the use of this particular program has accomplished under specified conditions.

Since the program manual will be designed for the use of teachers and school administrators, it should report the effectiveness of the program in straightforward terms. The reader should not be required to examine and interpret detailed technical data.

The manual should differentiate clearly and explicitly between (1) the opinions of experts or others about the effectiveness of the program, and (2) documented evidence on the outcomes actually obtained in practice. The manual should cite the sources of the available evidence to document any claims made about the effectiveness of the program. These sources should include not only those from the producer's own studies, but evidence from other carefully conducted studies as well. Many such independent studies, for example, may be carried out by school systems, research projects, or other agencies which have conducted program-assessment studies.

All claims or statements made in the manual about the effectiveness of a program (not directly supported in the manual) should be documented by specific reference to evidence contained in one or more detailed technical reports prepared in keeping with accepted standards of scientific writing. (Specific recommendations for the preparation of such documents are presented in Supplement II to this report.)

While brief summary statements concerning program effectiveness are appropriate in the program manual (as well as in advertising brochures, or on the cover or label of a program), such summary statements should always make reference to the technical report on which they were based. The manual should make explicit the correspondence between interpretive statements and the underlying data. The user should be told how and where he can obtain a copy of the technical report even though the manual is a derivative of such a report.

The manual should describe in a straightforward manner the studies undertaken to evaluate the program. The reader should be informed with respect to how the students were selected, the way in which the program was administered, the nature of the evaluative measures, and the results obtained. The description of these empirical studies should be nontechnical but precise. The data may be presented in simple form, such as through the use of graphs.

Any data on the program's effects on students' attitudes should be accompanied by a clear statement that student interest does not necessarily indicate program effectiveness. Where attitudes are measured solely through verbal responses, there should be no implication that other behavioral indices of motivation are also affected. For example, students may say that they liked the program enormously, but none of them may volunteer to receive further instruction.

5. Information concerning practicality of the program use

The user needs to be able to find out from the manual whether or not the program would be feasible for use by him in his local situation. The manual, therefore, can tell the reader whether the programs are reusable, and, if so, how many times they might reasonably be reused. Where supplemental material or equipment is needed, these should be described, along with statements of initial and maintenance costs. Information should be available to the reader about supervision requirements for the students taking the program. It is desirable to know the median and range of training time required to present the program effectively. The reports of empirical studies should be complete enough to provide information of this kind. In short, the user should be able to determine within reasonable limits the cost of instructing students by means of the program.

6. Procedures for introduction and use of the program

The user should be given a clear picture of the instructional conditions necessary for the success of the program. The manual should present a clear and detailed description of the recommended procedures for introducing students to the program and for administering the instructional and evaluative activities.

SUPPLEMENT II

to

RECOMMENDATIONS FOR REPORTING THE EFFECTIVENESS OF PROGRAMMED-INSTRUCTION MATERIALS[1]

prepared by the AERA-APA-DAVI Joint Committee
on Programmed Instruction and Teaching Machines

Recommendations for Preparation of Technical Reports

(Revised 31 October 1965)

I. INTRODUCTION

A. Purpose and Scope

This supplement contains further recommendations intended to serve as a guide for those who are preparing or reviewing technical documentation in support of statements about a program's performance characteristics--i.e., the outcomes that the program's use will demonstrably produce under specified conditions. Accordingly, these recommendations are concerned with what should be dealt with in technical reports of formal assessment studies in order that the general scientific criterion of reproducibility may be fulfilled.

For such studies, a detailed account of all experimental procedures and instruments used in assessment should be provided, giving the information needed for critical review of the study to determine whether summarized results and interpretations are warranted. The technical report on such detailed evaluation studies should permit a reviewer to assess the reported results in the light of the adequacy of criterion measures, description of test populations, description of conditions of experimental program use, and degree of correspondence between experimental conditions of testing and of those of intended application.

(Some of the considerations of experimental design and technical reporting dealt with here may not apply to informal program tryouts used only as guidance to the programmer in revising early versions of a program, nor to rough, preliminary screening tests made by a prospective user to help decide whether a published program seems to be generally suitable for his purposes.)

---

[1] Membership and sponsorship of the committee are indicated on page 1 of the July 1965 revision of the report which this supplement accompanies ("Recommendations for Reporting the Effectiveness of Programmed-Instruction Materials").

B. Audience

This supplement is primarily addressed to behavioral scientists or educational research workers who provide technical assistance to the program user or producer in obtaining or interpreting comprehensive assessment data. This includes both those who prepare technical reports on effects of programs, and those who advise purchasers concerning the soundness of reported data and statements concerning program effectiveness based on such data.

It is assumed that these individuals will have general competence in the fields of educational psychology, measurement, and experimental design. They should also be familiar with specialized technical considerations concerning program assessment studies, such as those discussed in papers by Lumsdaine[2,3], and by Jacobs, Maier, and Stolurow[4].

C. Main Considerations in Assessing Program Effects

Recommendations are given below for the four following elements in studies of program effects:

1. Criterion measures. Behavioral indices of what students can do after going through a program, including definition of potential outcomes and their exemplification in appropriate criterion tests or other measures reflecting attainment of these outcomes.

2. Characteristics of students. A complete description of the initial characteristics of the student population for which program assessment data is being reported.

3. Utilization procedures and experimental design. Procedures and arrangements for administering programs to defined samples of students under controlled and reproducible conditions, including procedures for administering the criterion tests.

4. Processing, analysis and reporting of data. Procedures for data processing and analysis, meeting scientific standards of reproducibility while also providing a basis for reporting the results in terms intelligible to the prospective program user.

---

[2] Lumsdaine, A. A. Some problems in assessing instructional programs. In Prospectives in Programming (R. Filep, Ed.). New York: Macmillan, 1963, pp. 228-62.

[3] Lumsdaine, A. A. Assessing the effectiveness of instructional programs. In Teaching Machines and Programmed Learning, II: Data and Directions (R. Glaser, Ed.). Washington, D.C.; National Education Association, 1966, pp. 267-320.

[4] Jacobs, P. I., Maier, M. H., and Stolurow, L. M. A Guide to Evaluating Self-Instructional Programs. New York: Holt, Rinehart and Winston (In Press).

II. TECHNICAL RECOMMENDATIONS

A. Criterion Measures

1. Description of tests. The report should give a detailed description of content areas and corresponding behaviorally stated outcomes that were tested as possible effects of the program. These would include any instructional outcomes which were measured as possible effects of the programmer.

   a. Standardized tests. If standardized, published tests are used, norms should be furnished or published sources of normative data should be cited. If in the context of reporting data in relation to such norms, data are also given for sub-samples of items for which separate norms have not been published, the rationale and methods for selection of items should be explained, and the recommendations set forth below for nonstandard tests should be followed.

   b. Nonstandardized tests. Since standardized tests will generally not suffice to provide detailed evaluation of specific strengths and weaknesses of a program, specially constructed tests will commonly be used for program evaluation studies. The considera tions set forth below apply particularly to such nonstandard tests.

   c. Where appropriate, the use of relevant behavior samples, not limited to paper and pencil tests or other verbal measures, is encouraged. The procedures used in obtaining any such behavioral measures, of course, should be fully described.

2. Detailed identification of test content. Copies of all test items used in measuring the possible outcomes that were tested should be included in the report if possible or, if not, should be available in an appendix. (Such appendix material as well as the basic report should be made permanently available from a suitable depository: University Microfilms, American Documentation Institute, etc., and the source of any such supplementary material should be referenced in the technical report.) Test items may be presented in the form of complete specimen copies of the tests used in assessing the program effects. However, the test's content should also be described as accurately as possible in overall terms, and the test items, insofar as possible, should be keyed to the statements of specific categories of outcomes. Such an analysis should show which items from the test were used to measure each kind of outcome. For example, within the subject matter of algebra, one might identify the test questions used to measure the student's ability to solve quadratic equations of a given type.

3. Scoring keys. The report (or appendix) should include not only copies of the test items themselves, but also a specification of the answers considered to be acceptable and unacceptable for each. Otherwise the procedure will not be reproducible, and it will be difficult for a reviewer to ascertain just what a percentage of "correct answers" really means.

4. Rationale for the construction of criterion measures.

   a. Definition of classes of outcomes. The rationale for the construction or selection of test items and other measures of program effects developed for use in an assessment study should be reported as fully as possible. Such a rationale should include as comprehensive as possible a characterization of the entire class of behaviors which were represented or sampled in each particular test or subtest (including any classes of items used for the purpose of measuring "transfer"). Such definitions should be clarified by giving examples.

   b. Sampling of items. To the extent possible, the report should explain in what way the particular samples of test items employed in the study were generated. (In other words, the report should show the basis for determining the extent to which a person who does well on a particular set of test items would also be likely to do well on any other sample of items generated in a similar manner.) For example, in a program in spelling, the report should not only specify the particular words included in the test, but also should describe the way in which the sample of test words was derived. As another example, various classes of quadratic equations might be identified, having specified ranges of coefficients and formats of expression with the samples of items used in the test(s) drawn from these according to a describable sampling plan.

   It is recognized that for some kinds of subject matter this kind of description must be quite imperfect at present, because of limitations in the state of the art of behavioral taxonomy. However, the report should give as complete a description as technically feasible at the present time so as to specify as accurately as possible the categories of outcomes that were tested.

5. Independence of samples of items in program and test. Many programs include only a relatively small sample of instructional items (frames) for a given objective; likewise, a feasible test of a program's effects will often contain only a relatively small sample of test items reflecting the kinds of outcomes to be assessed. In such cases, the report should indicate the procedures used to insure the independence of the sample of frames used in the program and the sample of items used in the criterion test (or tests). Specifically, it should show the extent of overlap and non-overlap of specific examples used in the program and in the test. Also, it should state what precautions were taken to insure that the program did not merely coach the student on a particular sample of items used in the test to assess its effects. Although it is desirable for the universe of possible test items, or samples thereof, to be known to the programmer, the particular sample used in a criterion test should be unknown to him at the time the program is written if the program is designed to teach behaviors that are supposed to generalize to other, similar items of behavior.

6. Measures of "transfer". If the "transfer" value of the program to other types of performances is reported, the rationale and methods for measuring transfer should be made explicit. As one example, an objective for a physics program might be to help students make new applications of principles not specifically dealt with in the program. Also, one might ascertain whether the program improved their ability to evaluate scientific experiments in other fields. The kind or degree of such "transfer" effects, if investigated, should be made explicit by description of the specific ways in which the "transfer" items differ from the content of the program.

   (If the program's objectives and content are comprehensive enough to cover all relevant behavioral outcomes, all of its effects could be considered direct effects, so that the concept of transfer would not be applicable. However, this conception does not fit the case in which instructional frames comprise only a partial sample of the total class of relevant items of behavior.)

7. Measures of interest and attitude. Where data are presented on the extent to which student "interests" or "attitudes" are influenced by the use of a program, the report should present copies of the instruments used, including all specific questions asked to assess interests or attitudes. It should also specify the conditions of administering these instruments (including such methodological precautions as anonymity of responses).

   Reports should make a clear distinction between data on students' interest in (or liking for) the program and gain in competence effected by the program. Their liking for the program itself should also be distinguished from interest aroused in the subject matter, and the student's liking for self-instructional programming generally should be distinguished from his liking of the particular program.

   Reports should also reflect a distinction between behavioral indices of motivation or interest engendered by a program, such as students volunteering to receive further instruction or being observed to engage in follow-up activities, versus mere verbal indicators believed to be predictive of such motivated behavior.

8. Effects of testing; use of parallel test forms. The report should indicate the way in which the study took into account possible spurious effects of the testing procedure, including those resulting from use of the same test more than once for a given subject.

   When parallel forms of measuring instruments are used (for example, the use of one form for a "before" test and one form for an "after" test, or the use of a parallel form in obtaining retention measures), their relationship to each other should be described fully. Any special techniques used (such as split forms with half of the group receiving one set of items before and the other half afterwards, while the reverse is true for another half of the group) should be explained in sufficient detail to be reproducible.

B. Characteristics of Students

1. Comprehensive description of relevant initial competences. The reader should be able to tell as precisely as possible what kinds of students were used in the study. They should be identified not only by relevant general background or prerequisite characteristics but also in terms of their initial status with respect to the competences to be developed by the program.

2. Detailed data on student characteristics. The report should identify in detail the characteristics of the students tested, including data on such factors as age, grade level, intelligence-test scores, reading ability, scholastic record and initial competence of the kinds measured as outcomes. Other factors which it might be important to report, depending on the program, might include visual and auditory acuity and any required special aptitudes such as manual dexterity. For such indices, appropriate measures of central tendency and spread (e.g., mean and standard deviation) should be supplied.

3. Expected vs. actual student competencies. The report should indicate any substantial discrepancy between the expected prior level of competence (as indicated in advertising, program manual, etc.) and the extent of actual relevant prior competences possessed by the experimental subjects.

4. Selection of subjects. The report should make clear how students were selected and assigned to the study. Reports should indicate how many students started and how many completed the program. For example, it should give relevant information about bases for potential selection bias, both in selection of schools or classes (e.g., a sample consisting only of those in which the teachers were willing to cooperate), and also in individual self-selection of typical students (e.g., the use of volunteers, bias due to dropouts). The characteristics of the dropouts should be reported in sufficient detail to determine the extent to which the remaining sample is representative.

   The report should state explicitly what was done to deal with the problem of dropouts or for absentees during the experiment. Measures for an attenuated subgroup should be accompanied by the earlier measures for that same subgroup, so that the data for two or more time points are both based on a common sample present at both time points. However, the number and characteristics of absentees left out of the sample should be clearly identified. Their pretest score when available should be reported in relation to the scores of those who finished so as to reveal any biases that may have resulted.

5. "Novelty effects". The report should state the extent of students' prior experience with programmed materials (and/or presentation devices) of the kind whose effects are being reported.

C. Conditions of Use and "Experimental Design"

The report should deal with two important aspects of "experimental design":

First, it should describe the technical procedures and controls employed, so that the reader can assess the extent to which reported gains in knowledge, skills, etc., can be validly defended as results of the program itself, rather than of other concurrent or prior sources of influence.

Second, it should specify the conditions of use of the program which affect the applicability or generalizability of the results. For example, it should describe conditions of utilization in a classroom, or the use of a program for individual study, whether students were required to complete all of the frames or whether they merely had the program materials available to them to proceed as far with as they chose, etc.

1. Generally applicable features. To serve these purposes effectively, the report should deal with such details as the following:

   a. The edition of the program used. The extent to which the program used to collect data was different, if at all, from the commercially available edition. (If more than one edition is available, the report should specify which was used.)

   b. Utilization situation. The kind of situation under which the program was administered (for example, used in regular classrooms or in special settings). The conditions of the program's use should be reported in sufficient detail so that their essential features could be reproduced by another investigator.

   c. Time intervals. The distribution of amount of time per day students spent on the program, for how long the program's use was continued, and time intervals between instruction and testing as well as between a first test and a later retention test. If not constant for all students, the distribution of such intervals should be given.

   d. External help. Any assistance supplied by the teacher or by others during the administration of the program, or at any time between the obtaining of pre- and post- (including retention) measures, should be fully reported. If teachers, or proctors, answered questions about the instructional content or procedures, full details should be given concerning control groups used to assess the effects of external assistance (see also considerations applying to the use of control groups given below under "Comparative Studies").

   e. Motivational conditions. The extent to which motivational influences were exercised, such as whether the teacher checked on students to make sure they were working on the programs, or whether they were left to work by themselves, whether students were tested at intervals (and if so, what tests were used and whether students told that the tests counted on their grades), whether students were given any other special incentives for working on their programs, or any disciplinary action for not working on it.

f. Testing conditions. Conditions under which criterion tests were given, including: (1) total time taken for the tests (including distribution of times if variable) and whether the students were held to announced time limits; (2) instructions given to the students about the tests; (3) precautions taken to avoid inappropriate help from teachers or other sources; and (4) any special incentives given in connection with testing.

g. Use of repeated measurements. The report should indicate whether the same subjects were used for "before" measures as for immediate testing after the completion of the program and for retention data.

h. Need for preprogram measures. If the experimenter has dispensed with a "before" measure (or equivalent measure from a separate uninstructed group) on the supposition that the student's initial level of knowledge is substantially zero, the basis on which this belief may be defended should be explicitly stated. In any case, the levels of attainment reported should in such instances be identified as measured competence following the program, rather than as gain or effects due to the program.

2. Comparative studies. The following additional considerations apply to any studies in which data for two or more different treatment groups are compared.

a. Purpose of comparison. The purpose of using comparison groups should be indicated--e.g., groups assigned to alternative programs or to alternative procedures for using a given program, or groups used as a control for extraneous sources of influence.

b. Definition of comparison treatments. When the effectiveness of a program is being compared with the effectiveness of some other instructional procedure, full reporting of the nature of the "other" instruction, such as to make it substantially reproducible, is essential for valid interpretation.

c. Assignment of subjects to treatments. The report should specify the procedures used to assign subjects to experimental treatments, e.g., by purely random assignment or random assignment of matched individuals.

d. Equivalence of groups. Reports on studies in which equivalent subgroups are used to obtain data at different time intervals should identify clearly the basis on which the comparability of these groups was established. Also, relevant activities intervening should be reported and any interaction between the groups should be noted.

e. Control for confounded factors. In any study in which the effects of one program or procedure are compared with those of another, procedural controls for insuring comparability of

conditions for the two treatments should be reported in full, together with any known factors that might impair such comparability.

3. Time vs. criterion achievement. A special problem in the assessment of self-instructional programs lies in the fact that there are two dependent variables of interest: (a) time spent in instruction; and (b) proficiency, e.g., gain in achievement level. In the comparison of two programs, it is possible for one to produce higher achievement scores than the other but also to require more time.

   Gain in achievement level is sometimes expressed as an "efficiency" ratio of gain divided by time. Any such derived measure should be clearly explained, so that such values as "percent efficiency" are not presented without the reader's being able to tell precisely what kind of derived measure was, in fact, employed. If no single achievement-time index seems defensible as a single figure of merit for a program's instructional efficiency, the alternatives for reporting are:

   a. Report gains in attainment of outcomes achieved or final levels of proficiency achieved by going through the program from beginning to end, separately reporting time spent on the program as a second dependent variable.

   b. Hold time constant experimentally, reporting attainment achieved in some arbitrarily fixed period of time, but preferably after two or three periods of time.

   c. Determine and report, as the main dependent variable, time required to achieve specified levels of attainment.

   The third alternative presumes that all students reach some minimum level of proficiency. This involves repeated testing of each student's progress. Time-to-criterion can be employed as a sole dependent measure only if the basis for determining when the student has achieved criterion is based on such successive testing. Since the null hypothesis cannot be proved, it is not sufficient merely to show that two groups who took different amounts of time to complete alternative programs "did not differ significantly" with respect to the criterion level attained. Such statements should be scrupulously avoided.

D. Analysis and Reporting of Data

   1. General considerations. Results should be reported for all effects of the program which the study attempted to measure, including possible effects outside the primary objectives of the program and regardless of whether significant gains from the program were shown by the data.

2. Analysis of specific program effects. Data for tests given before and after students have taken the program should be given for total scores and for content subscores, so that a differential profile of program effectiveness can be made. For these measures, in fine, the report should present summary statistics such as means and standard deviations. Such data should be given not only for the total group of subjects, but also for subgroups differentiated by relevant student characteristics such as ability and initial knowledge. Such analyses should be accompanied by appropriate statistical tests of the reliability of any differential effects reported.

3. Tests of significance and confidence limits. Enough information should be supplied to allow the reader of the technical report to check on the appropriateness of any inferential statistics reported --i.e., tests of significance or fiducial limits. Where the differences between two sets of scores are not statistically significant, the report should avoid the error of saying that the two sets of scores were the "same" or "equivalent" results. Confidence limits for percentage values should be indicated when reporting for individual items and for means or other average score measures. The method by which confidence limits or significance tests are computed should be reported explicitly, since practice in computing such statistics is not uniform, so that an evaluator may verify the computation.

4. Derived measures. The use of "percentage gain" or "percentage retention" measures, particularly when unqualified, is discouraged. Such measures should be accompanied by the basic data from which they are derived, and by an explicit indication of how the percentage measure was obtained. They should also be accompanied by an indication of the standard error of such measures or by the data from which these standard errors can be derived.

5. Reporting of basic data. Any relevant details of assessment data which require more space than is appropriate for a published report should be made available (for example, in a supplementary report or by deposit with such agencies as The American Documentation Institute University Microfilms). For example, supplementary tables should, whenever possible, be provided for the technical report, showing the complete matrix of all individual subjects' responses to all individual test items.

   a. Such an N x k matrix (N = no. of Ss, k = no. of items) should be deposited with such an agency as ADI, to permit checking and re-analysis of the data as desired by the technical reviewer.

   b. This information should be accompanied by each of the available scores of prerequisite knowledge and ability of each subject, so that analysis can be made of results for ability subgroups other than those employed by the original report of the assessment study.

III. CHECKLISTS

The checklists below are intended to recapitulate main points of the recommendations made in this supplement. In a well-reported assessment study it should be possible to answer all of these questions in the affirmative. The information needed as a basis for such answers is indicated more fully by the recommendations given in Sections II - A, B, C and D, above. Affirmative answers to all questions do not guarantee the validity of the results of a study, but negative answers to any of the questions may call the validity of the study into question.

CHECKLIST "A": Criterion Measures

1. Does the report clearly identify the test instruments used to measure the behavioral effects of the program that were tested?

2. Are specimen copies of the tests and other measures supplied in the report or in an appendix?

3. Does the report supply and interpret the scoring key for all items?

4. Is the rationale for test content clearly specified both in terms of behavioral categories and also by showing how the particular test items used to exemplify each category were generated?

5. Are there adequate safeguards against spurious effects due to selective coaching by the instructional program on specific items of the criterion tests?

6. Is the evidence clearly presented to indicate the nature of any effects reported for "transfer" to types of behavior not directly dealt with in the program?

7. Are instruments used to measure interest or attitudes provided, and are the conditions affecting their validity adequately described? Does the report clearly distinguish among effects on achievement and on interest or motivational effects?

8. Does the report deal adequately with special conditions affecting validity of measurement, including use of parallel test forms?

CHECKLIST "B": Characteristics of Subjects

1. Has the report described clearly and completely the kind of student population with which the program was used which might influence the effectiveness of a particular program?

2. Similarly, has the report indicated what students were able to do, with respect to the outcomes tested, before they started the program?

3. Are the characteristics of test population and intended population substantially the same?

4. Has the report described how the schools and students were selected for the study so as to identify possible sources of selection bias, and does it deal adequately with such sources of selection bias, including bias due to dropouts?

5. Has the extent of student's prior experience with programs been taken into account?

CHECKLIST "C": Conditions of Use and Experimental Design

1. Conditions of reproducibility of program administration. Does the report supply complete information regarding the way the program was used so that it could be administered again in the same way?

   a. Does the report indicate the form or edition of program that was used?

   b. Are the conditions under which the program was presented fully described?

   c. Are the time periods for program use and testing specified fully?

   d. Does the report describe fully the kind and amount of assistance supplied to students in the use of the programs?

   e. Are motivational conditions affecting students' work on the program adequately described?

   f. Does the report describe the conditions under which the tests were given, including use of repeated measures and any conditions which might alter the validity of the testing?

   g. Are numbers of cases tested fully reported, including identification of any dropouts, and are measures for different time points based on equivalent samples of students?

2. Validity of comparative studies

   a. Where comparative results are given for alternative treatments, does the report describe the nature of the alternative treatments in a way that permits reproducibility?

   b. Does the report show that the students in the different comparison groups are equivalent samples, and does it describe the precise method of assigning students to alternative treatment groups?

CHECKLIST "D": Analysis and Reporting

1. Does the report give completely and usefully the results of pre- and posttests and other evaluative measures?

a. Are the results given not only for the total test, but also for the subtests which indicate specific outcomes attained more or less effectively by the program?

b. Are the test scores presented for each of the main subgroups in the student sample, and are the subgroups defined by relevant characteristics such as ability and background?

2. Are the methods for computing fiducial limits and tests of significance available so they can be verified?

3. Are all derived measures such as "percentage retention" and "percentage gain" clearly explained, and are the basic data on which they are based reported?